Assignment No. 1

LP2- ETL MODEL

1.1 Title:

For an organization of your choice, choose a set of business processes. Design star / snow flake schemas for analyzing these processes. Create a fact constellation schema by combining them. Extract data from different data sources, apply suitable transformations and load into destination tables using an ETL tool.

1.2 Problem Definition:

Design a basic ETL model using Rapid Miner Application.

1.3 Prerequisite:

- Basic concepts of ETL.
- Knowledge about Rapid miner tool.

1.4 Software Requirements:

- Rapid Miner

1.5 Hardware Requirement:

- PIV, 2GB RAM, 500 GB HDD, Lenovo A13-4089Model.

1.6 Learning Objectives:

Understand the implementation of the various ETL model using Rapid Miner tool.

1.7 Outcomes:

After completion of this assignment students can develop and analyze the ETL model and will understand the working.

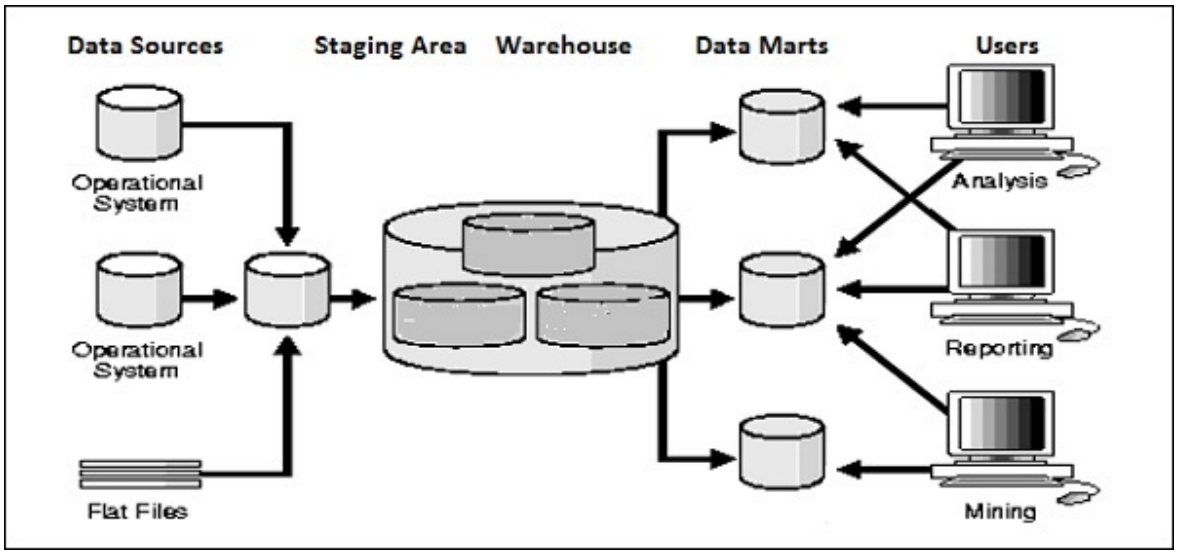
1.8 Theory Concepts:

What does ETL mean?

ETL stands for Extract, Transform and Load. An ETL tool extracts the data from different RDBMS source systems, transforms the data like applying calculations, concatenate, etc. and then load the data to Data Warehouse system. The data is loaded in the DW system in the form of dimension and fact tables.

Extraction

- A staging area is required during ETL load. There are various reasons why staging area is required.
- The source systems are only available for specific period of time to extract data. This period of time is less than the total data-load time. Therefore, staging area allows you to extract the data from the source system and keeps it in the staging area before the time slot ends.
- Staging area is required when you want to get the data from multiple data sources together or if you want to join two or more systems together. For example, you will not be able to perform a SQL query joining two tables from two physically different databases.
- Data extractions' time slot for different systems vary as per the time zone and operational hours.
- Data extracted from source systems can be used in multiple data warehouse system, Operation Data stores, etc.
- ETL allows you to perform complex transformations and requires extra area to store the data.



Transform

In data transformation, you apply a set of functions on extracted data to load it into the target system. Data, which does not require any transformation is known as direct move or pass through data.

You can apply different transformations on extracted data from the source system. For example, you can perform customized calculations. If you want sum-of-sales revenue and this is not in database, you can apply the **SUM** formula during transformation and load the data.

For example, if you have the first name and the last name in a table in different columns, you can use concatenate before loading.

Load

During Load phase, data is loaded into the end-target system and it can be a flat file or a Data Warehouse system.

1.8.1 Tool for ETL: *RAPID MINER*

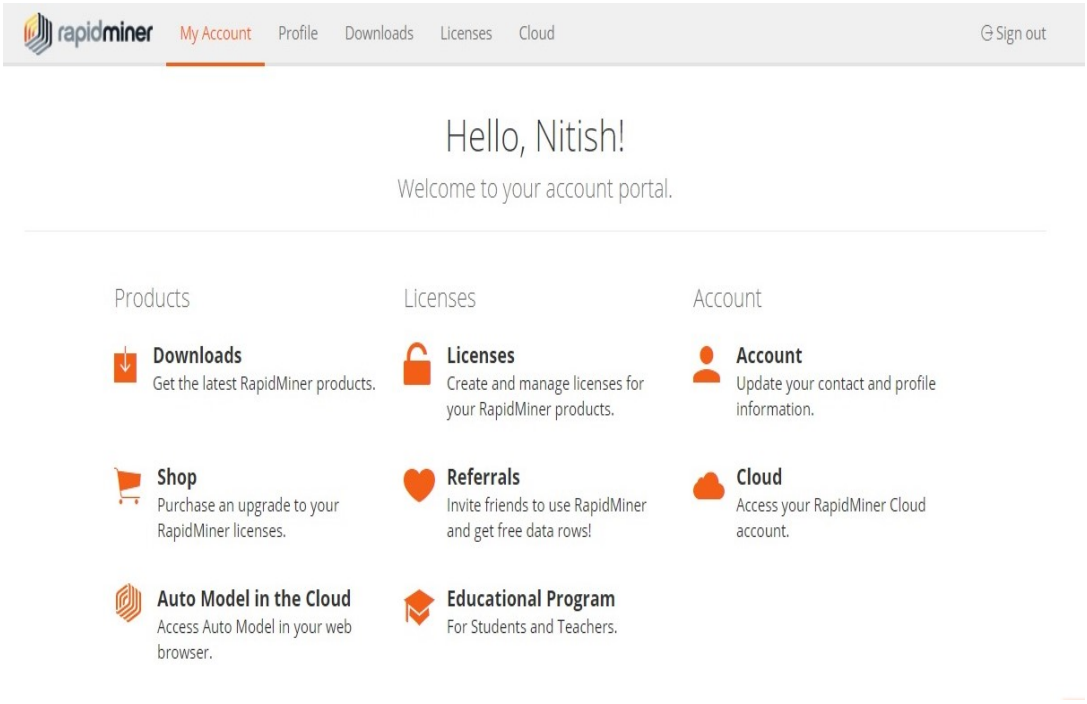
Rapid Miner is a world-leading open-source system for data mining. It is available as a stand-alone application for data analysis and as a data mining engine for the integration into own products. **Rapid Miner is now Rapid Miner Studio** and Rapid Analytics is now called Rapid Miner Server.

In a few words, Rapid Miner Studio is a "downloadable GUI for machine learning, data mining, text mining, predictive analytics and business analytics". It can also be used (for most purposes) in batch mode (command line mode)

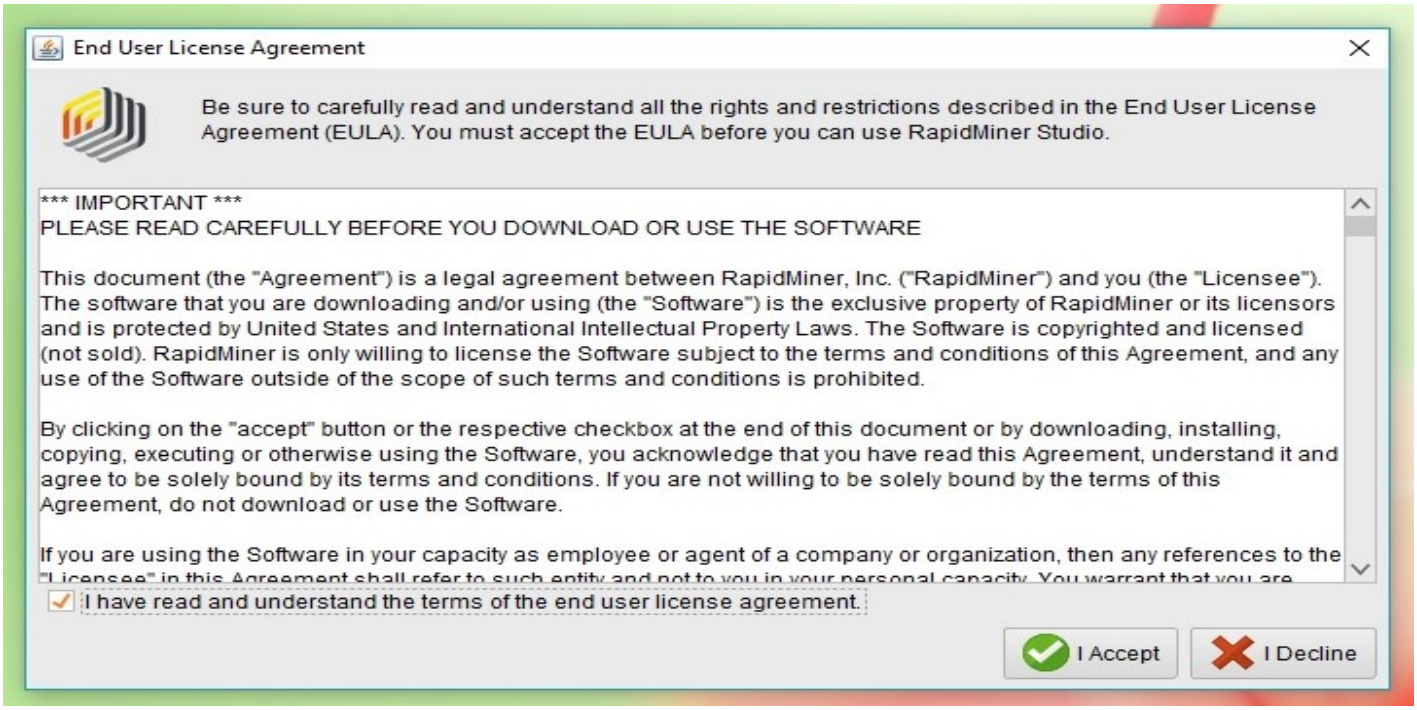
Rapid Miner Support to Nominal, Numerical values, Integers, Real numbers, 2-value nominal, multi-value nominal etc.

STEPS FOR INSTALLATION:

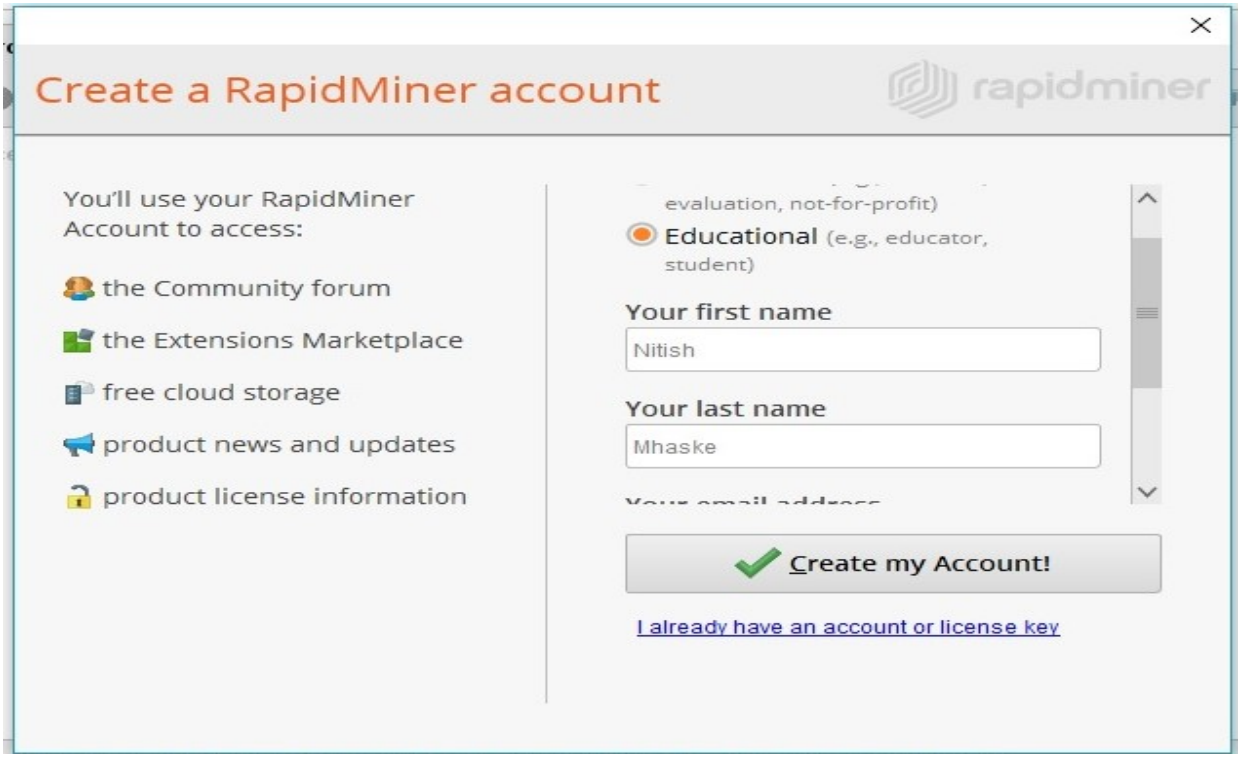
1. Downloading Rapid Miner



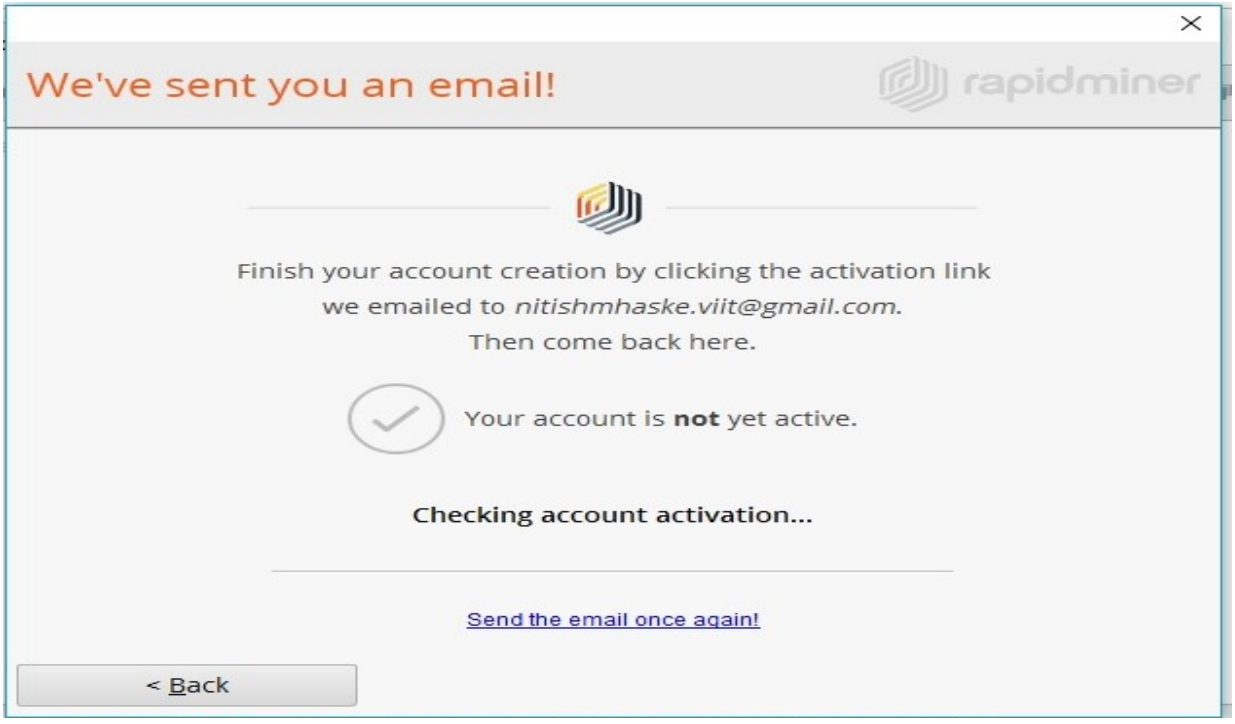
2. Installing Rapid Miner



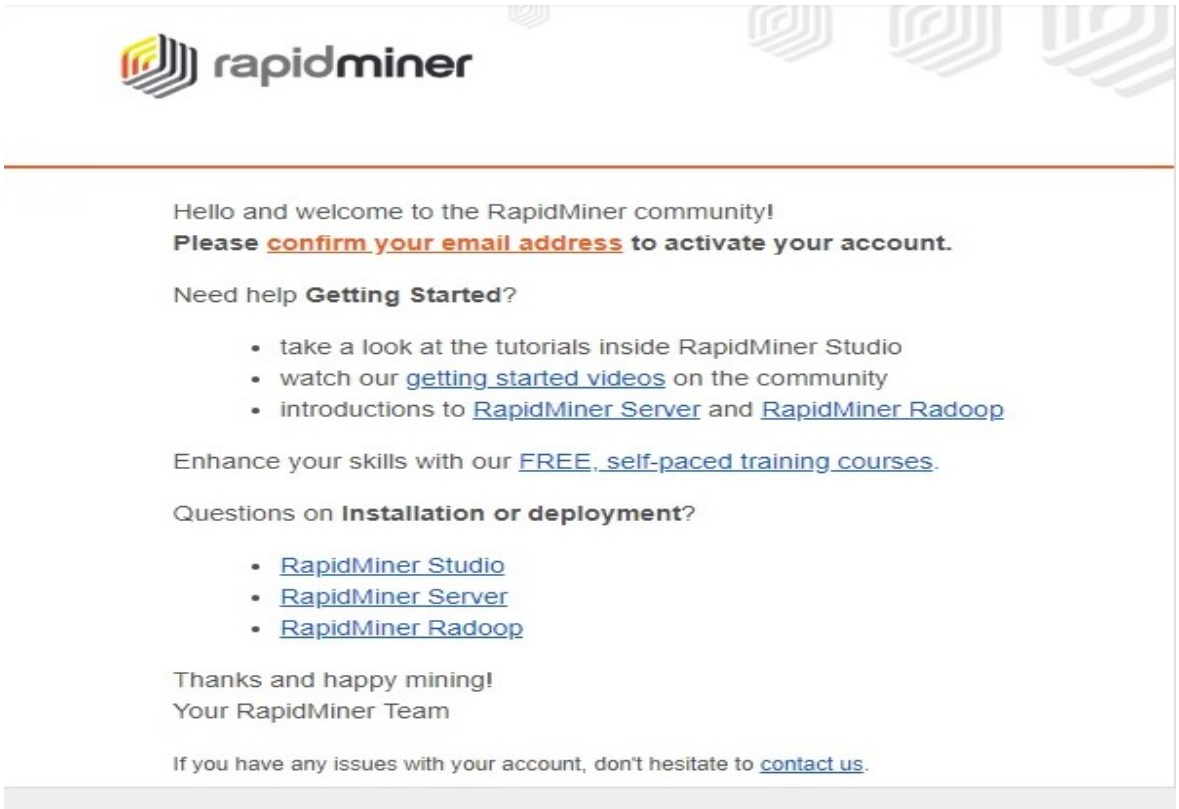
3. Configuring Rapid Miner Server settings



4. Configuring Rapid Miner Server's Account



5. Confirm the Rapid Miner Account



6. Completing the installation

Data Warehousing Schemas

1. Star Schema
2. Snowflake Schema
3. Fact Constellation

Star Schema

For example, as you can see in the above-given image that fact table is at the center which contains keys to every dimension table like Deal_ID, Model ID, Date_ID, Product_ID, Branch_ID & other attributes like Units sold and revenue.

Characteristics of Star Schema:

- Every dimension in a star schema is represented with the only one-dimension table.
- The dimension table should contain the set of attributes.
- The dimension table is joined to the fact table using a foreign key
- The dimension table are not joined to each other
- Fact table would contain key and measure
- The Star schema is easy to understand and provides optimal disk usage.
- The dimension tables are not normalized. For instance, in the above figure, Country_ID does not have Country lookup table as an OLTP design would have.
- The schema is widely supported by BI Tools

Snowflake Schema

A Snowflake Schema is an extension of a Star Schema, and it adds additional dimensions. It is called snowflake because its diagram resembles a Snowflake.

The dimension tables are normalized which splits data into additional tables. In the following example, Country is further normalized into an individual table.

Characteristics of Snowflake Schema:

- The main benefit of the snowflake schema it uses smaller disk space.
- Easier to implement a dimension is added to the Schema
- Due to multiple tables query performance is reduced
- The primary challenge that you will face while using the snowflake Schema is that you need to perform more maintenance efforts because of the more lookup tables.

Star Schema	Snow Flake Schema
Hierarchies for the dimensions are stored in the dimensional table.	Hierarchies are divided into separate tables.
It contains a fact table surrounded by dimension tables.	One fact table surrounded by dimension table which are in turn surrounded by dimension table
In a star schema, only single join creates the relationship between the fact table and any dimension tables.	A snowflake schema requires many joins to fetch the data.
Simple DB Design.	Very Complex DB Design.
Denormalized DS & query run faster.	Normalized Data Structure.
High level of Data redundancy	Very low-level data redundancy
Single Dimension table contains aggregated data.	Data Split into different Dimension Tables.
Cube processing is faster.	Cube processing might be slow because of the complex join
Offers higher performing queries using Star Join Query Optimization. Tables may be connected with multiple dimensions.	The Snow Flake Schema is represented by centralized fact table which unlikely connected with multiple dimensions.

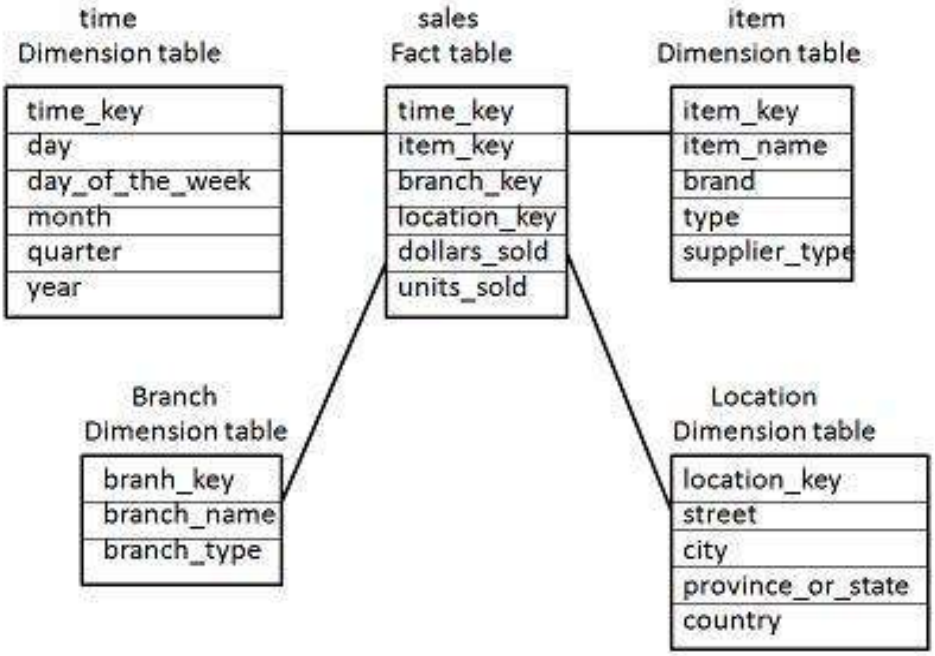


Diagram : Star Schema

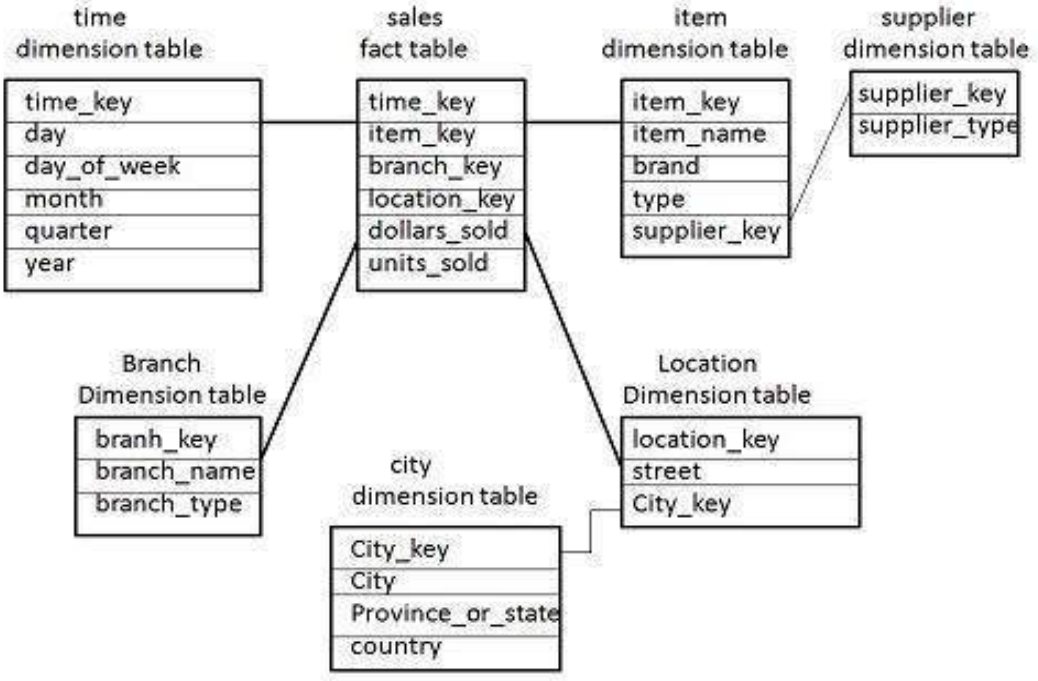
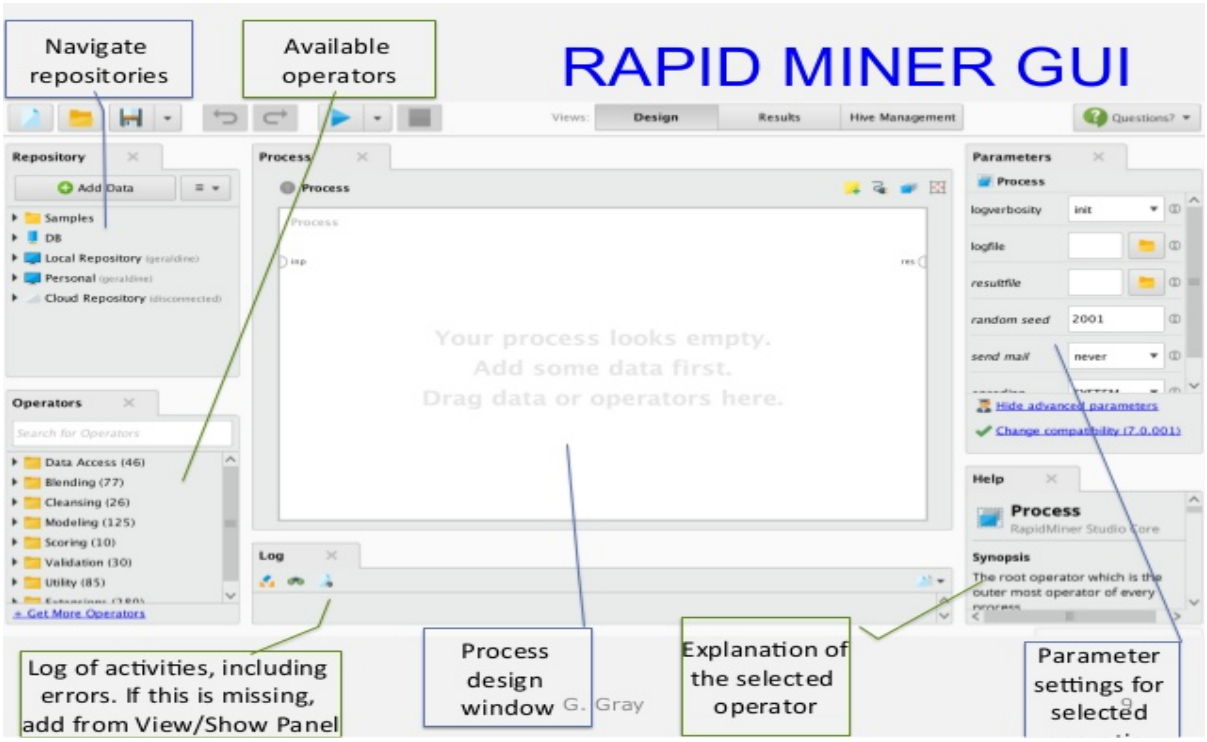
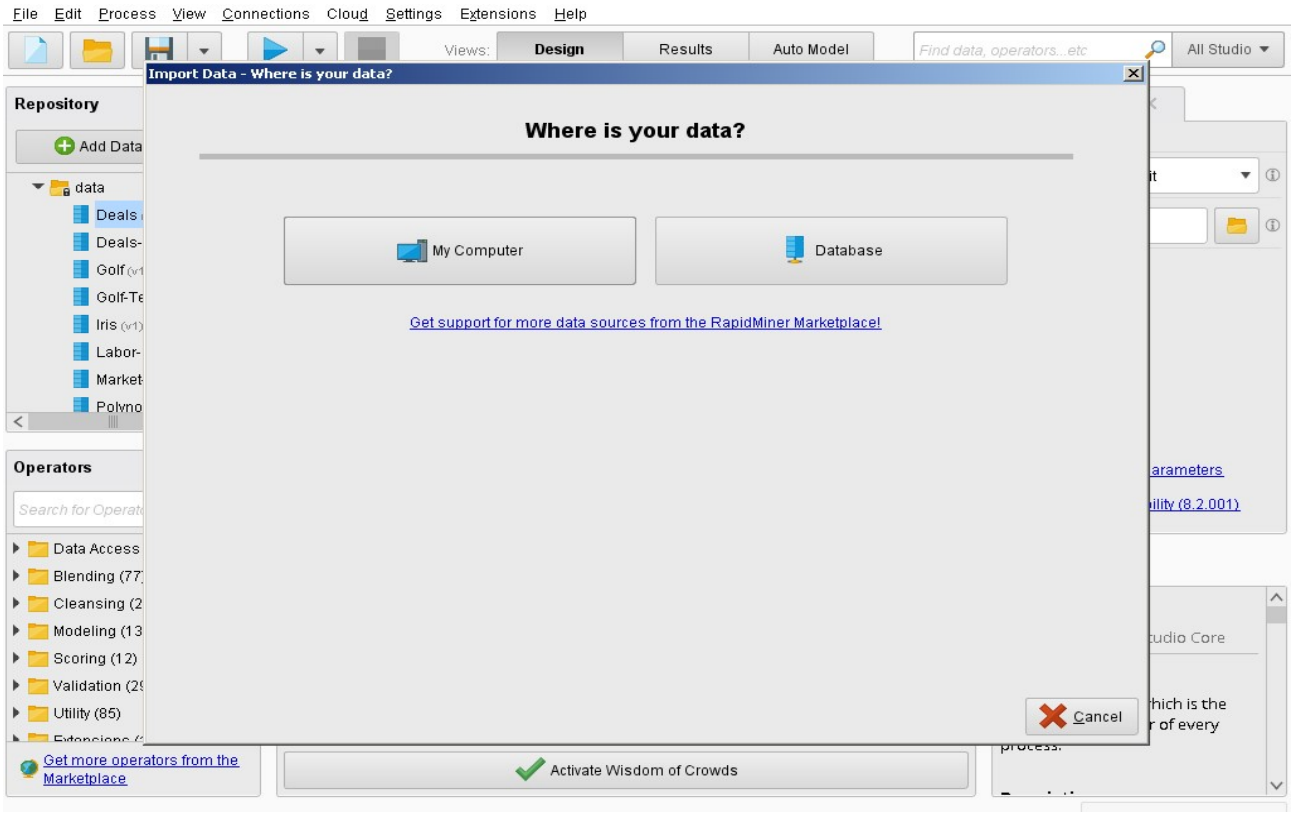


Diagram : Snowflake Schema

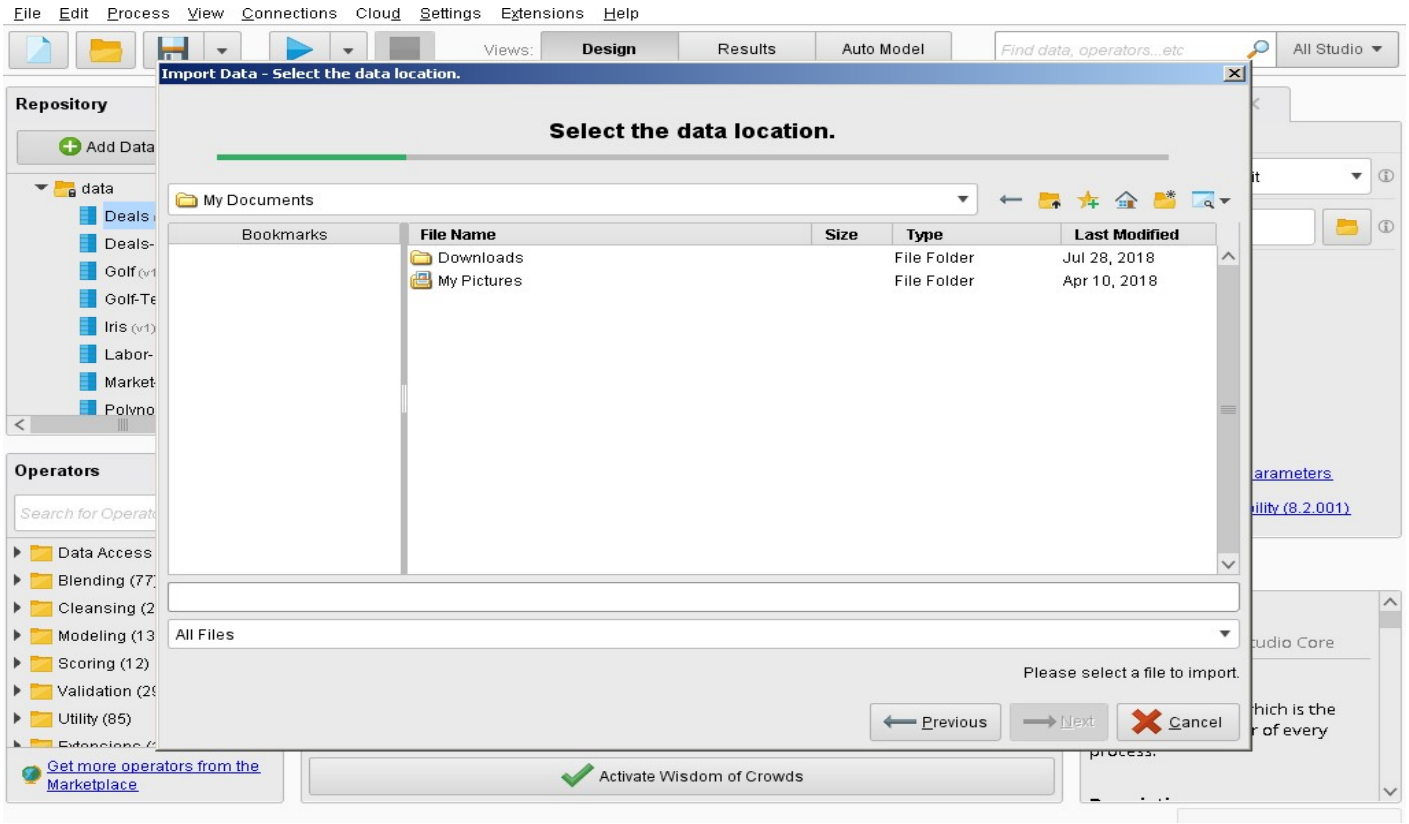
1. Design Model



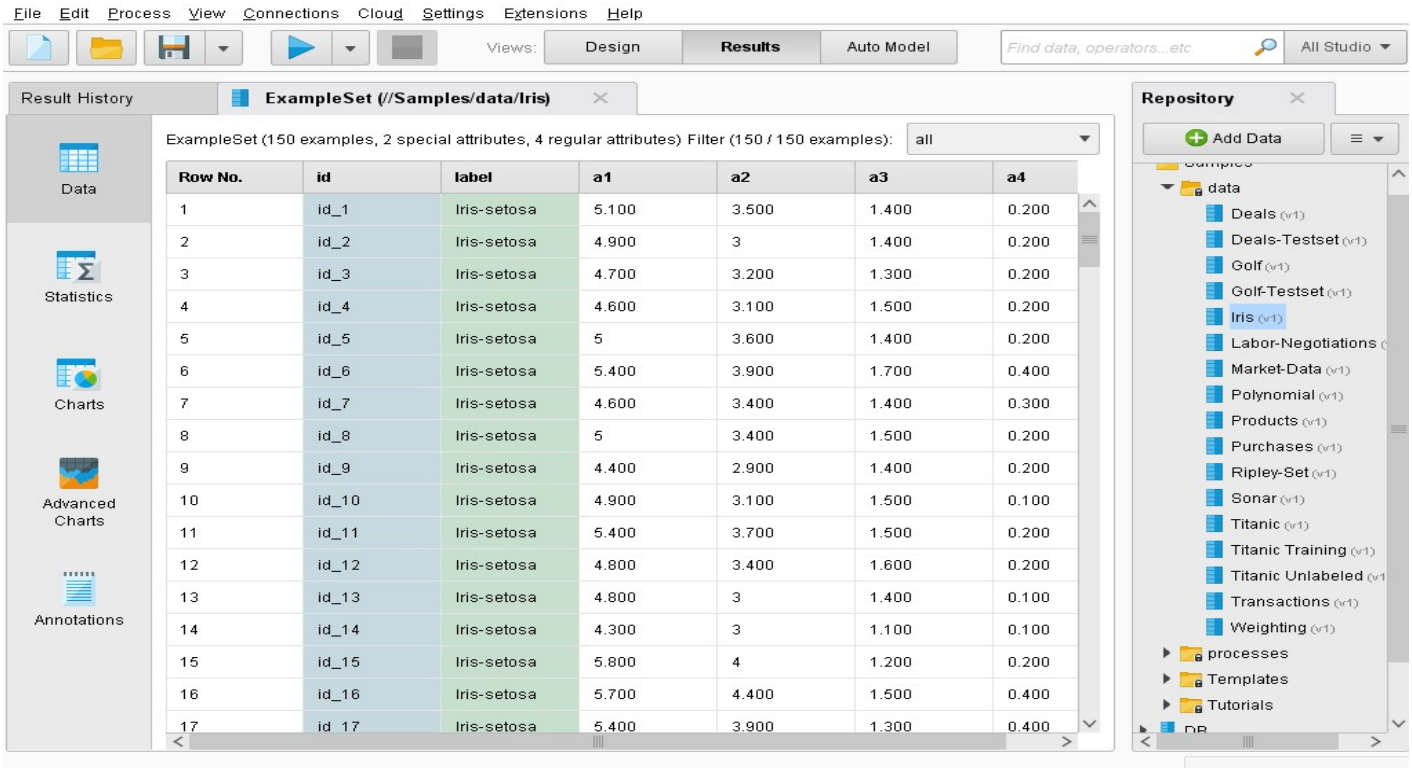
Step-1 Import Data from Source



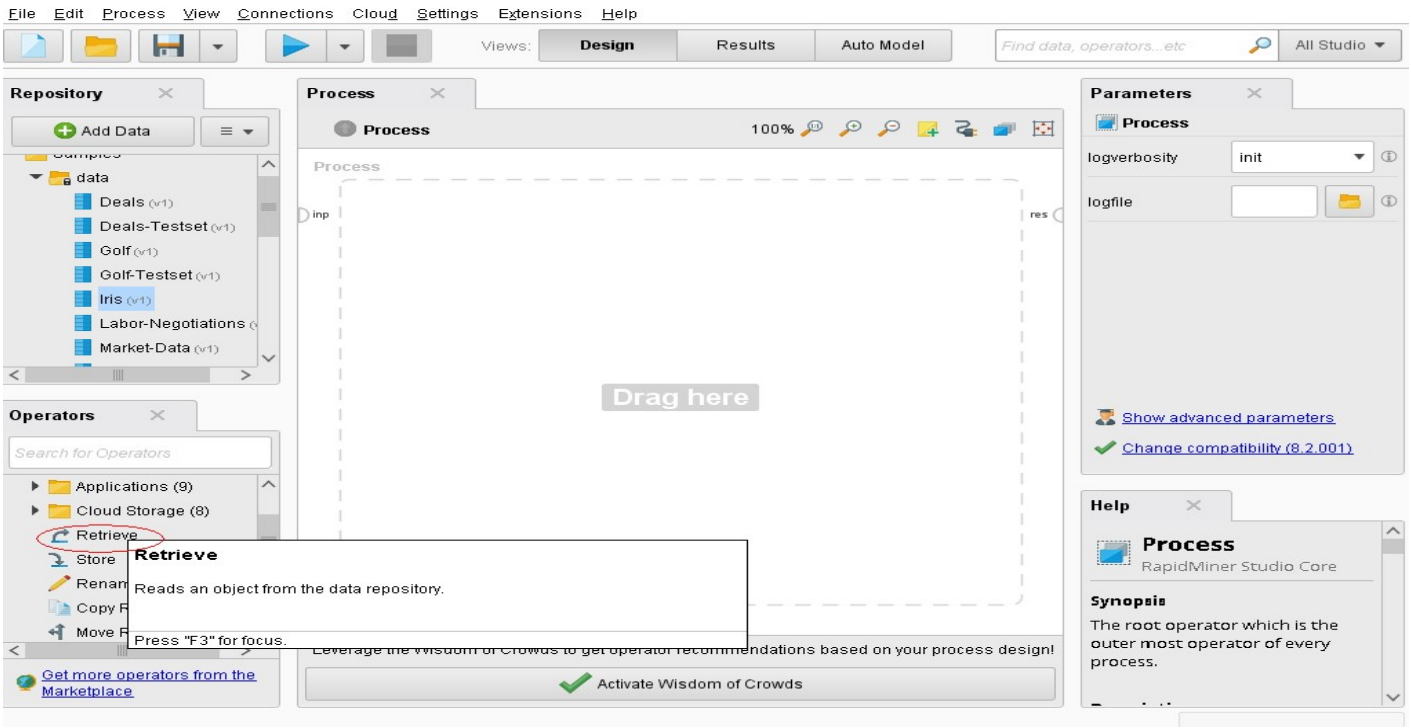
Step-2 Select Data Location



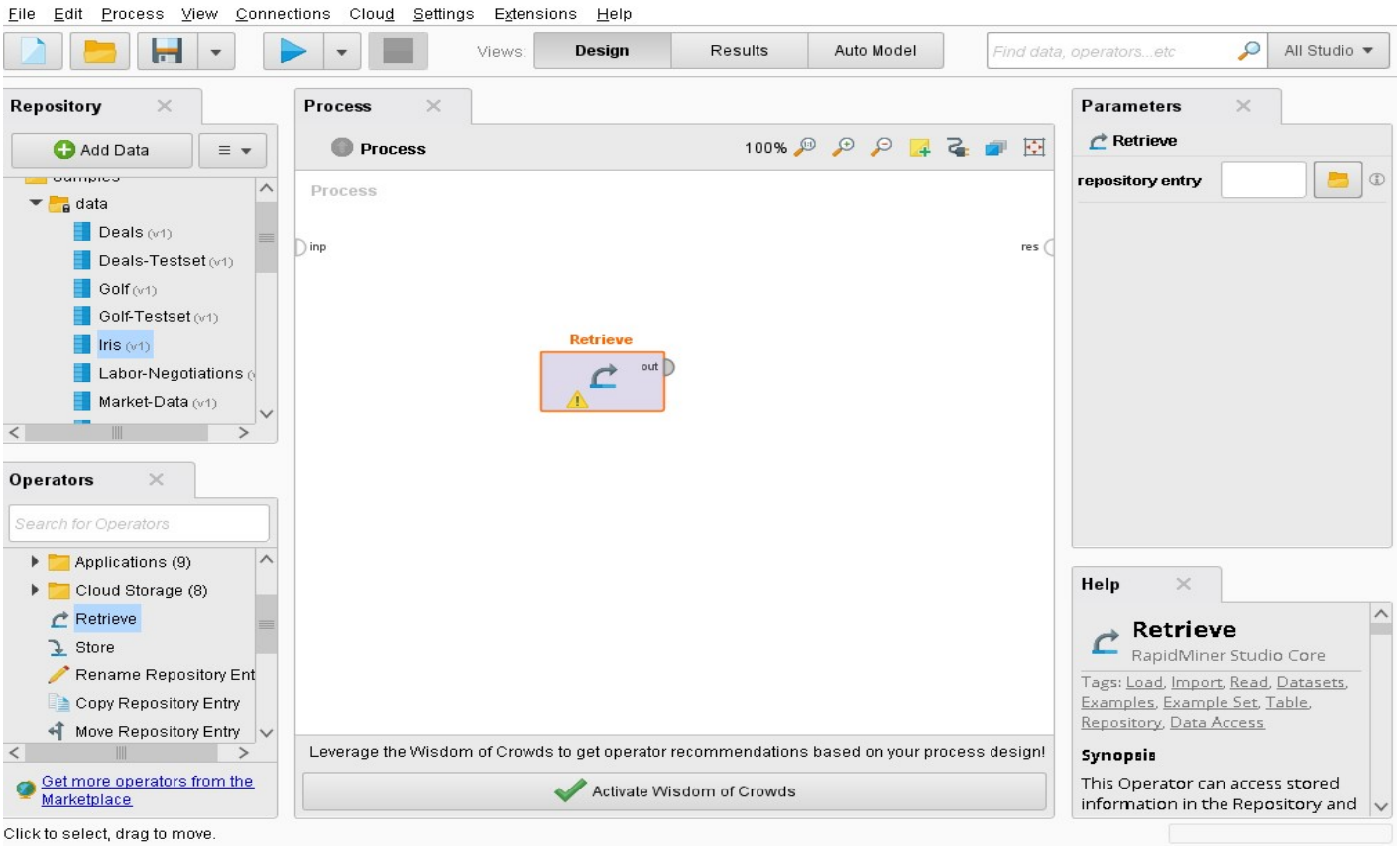
Step-3 Open Sample Data Set eg. Iris dataset available inbuilt with tool



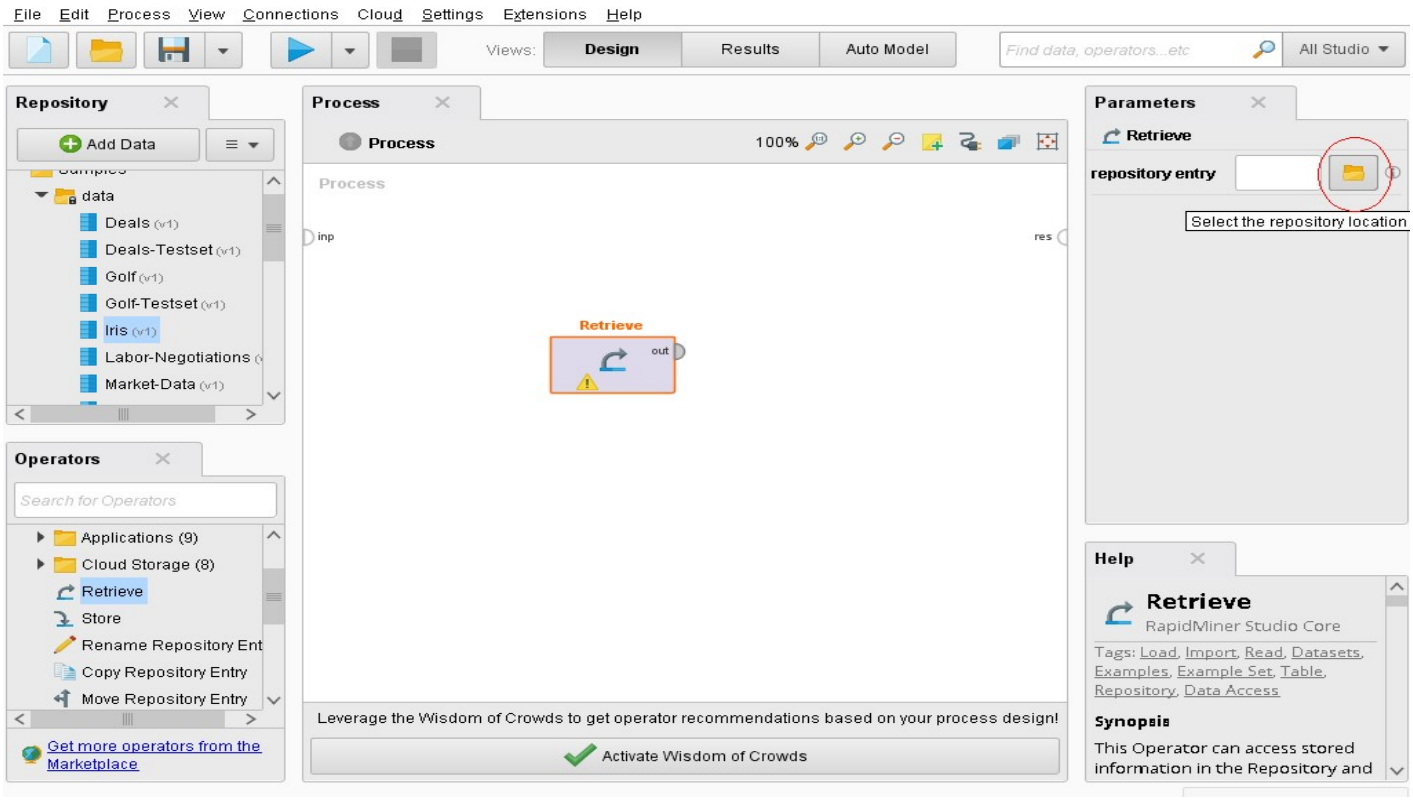
Step-4 Click on retrieve Operator Drag in Process View



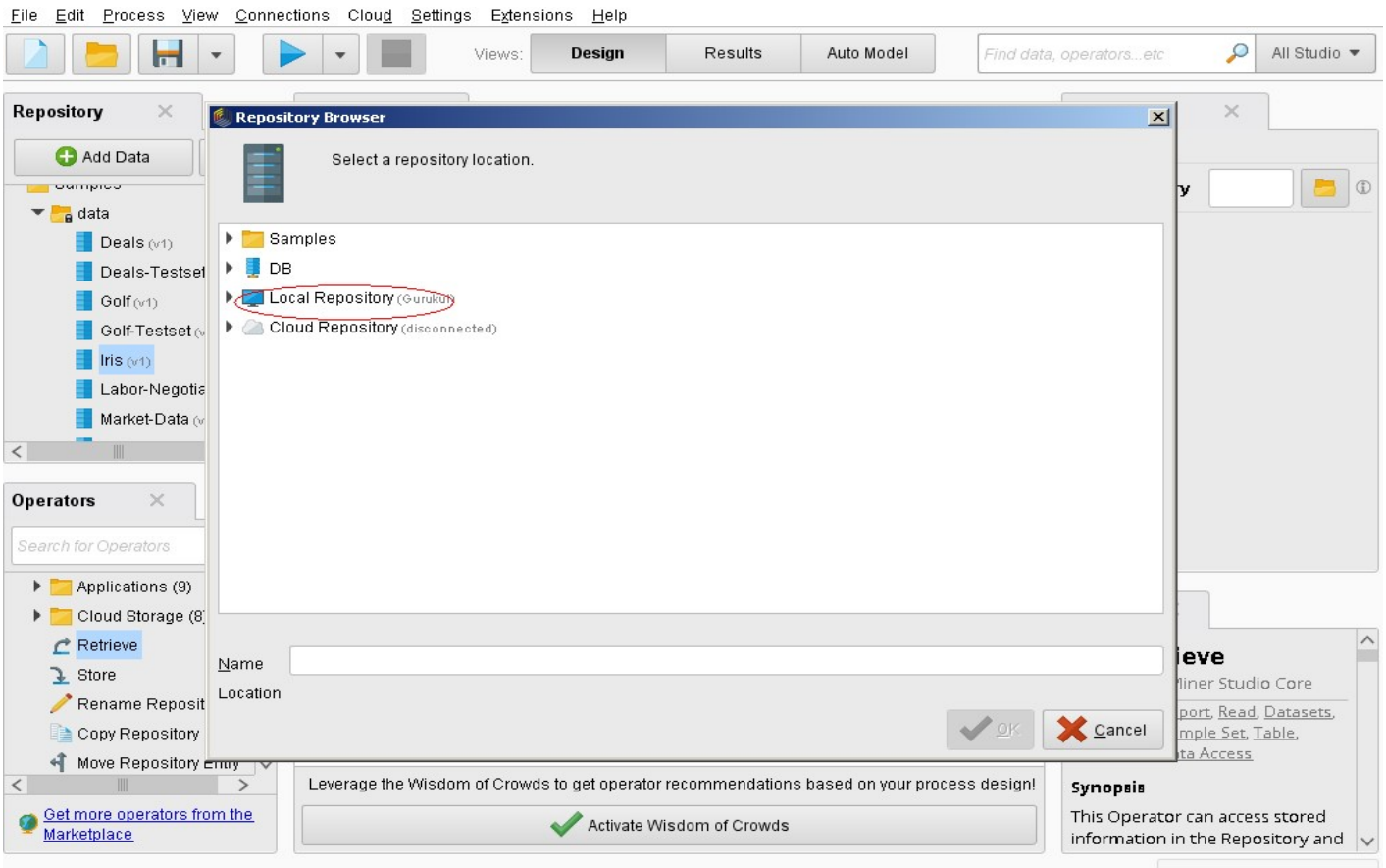
Step-5 Retrieve icon shows in Process View it has input and out Operator



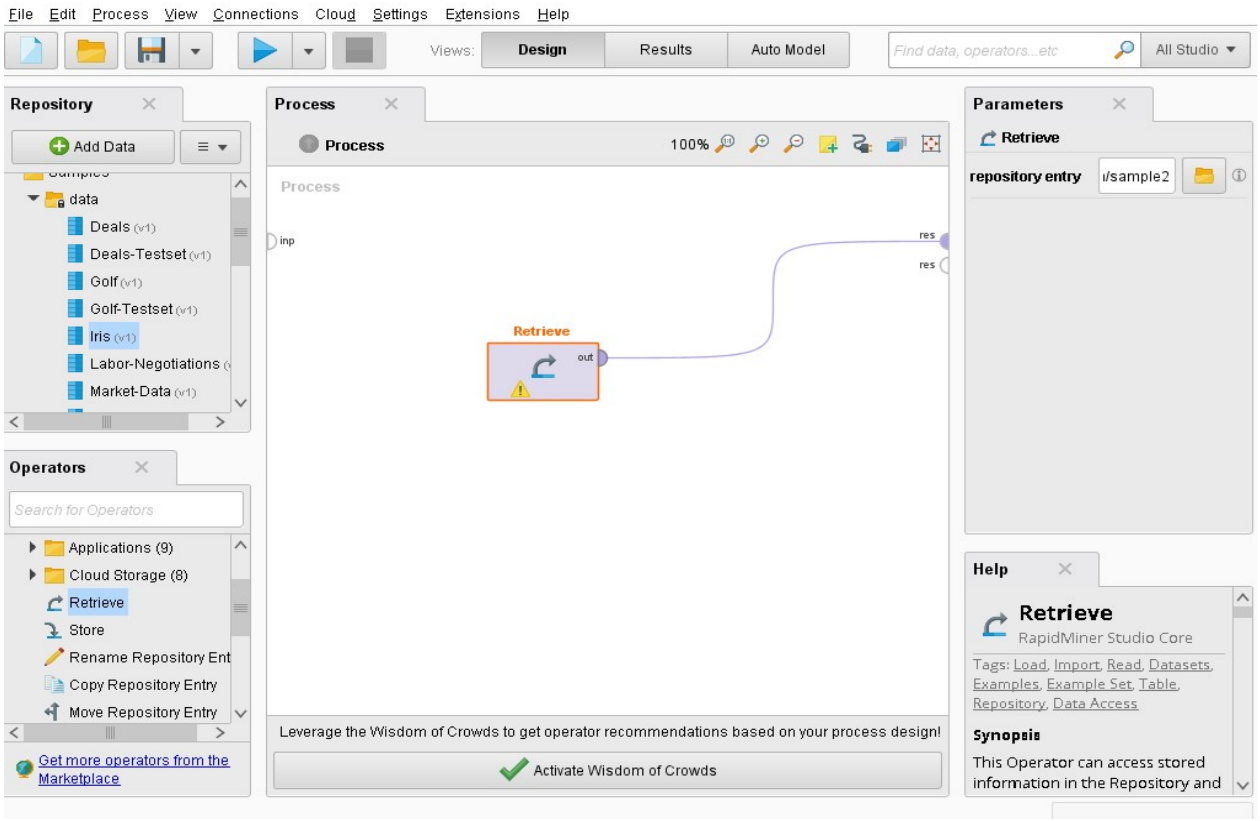
Step-6 Click on repository entry



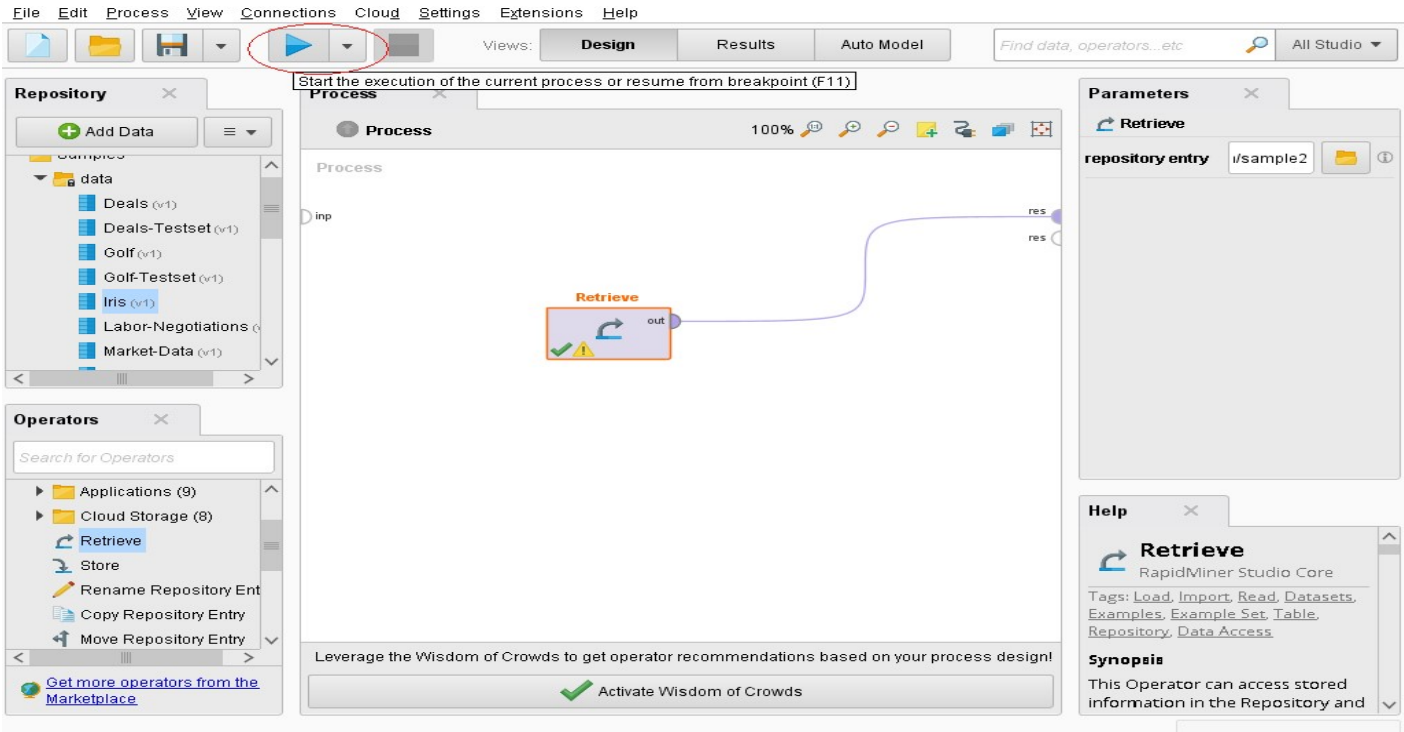
Step-7 Select Local Repository



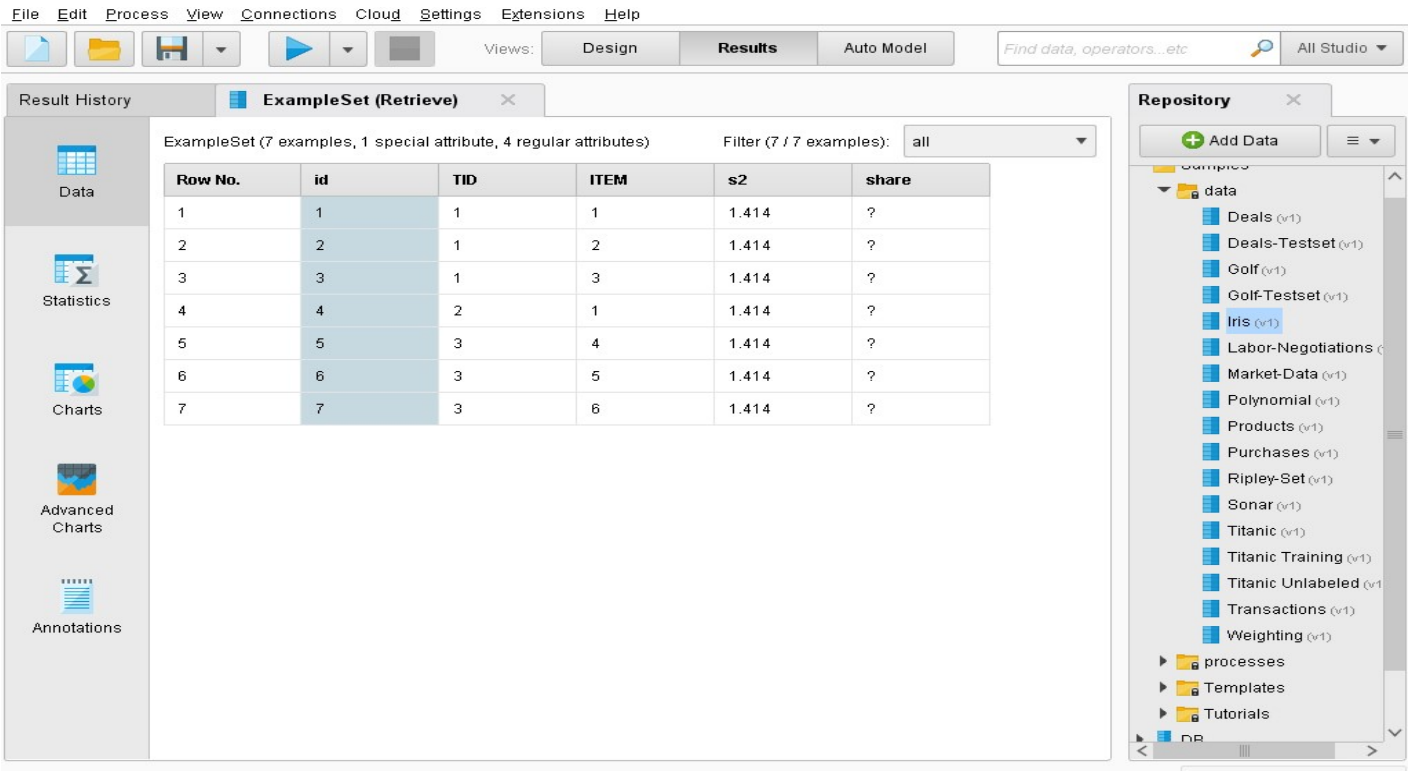
Step-9 Join Out Operator to result Operator



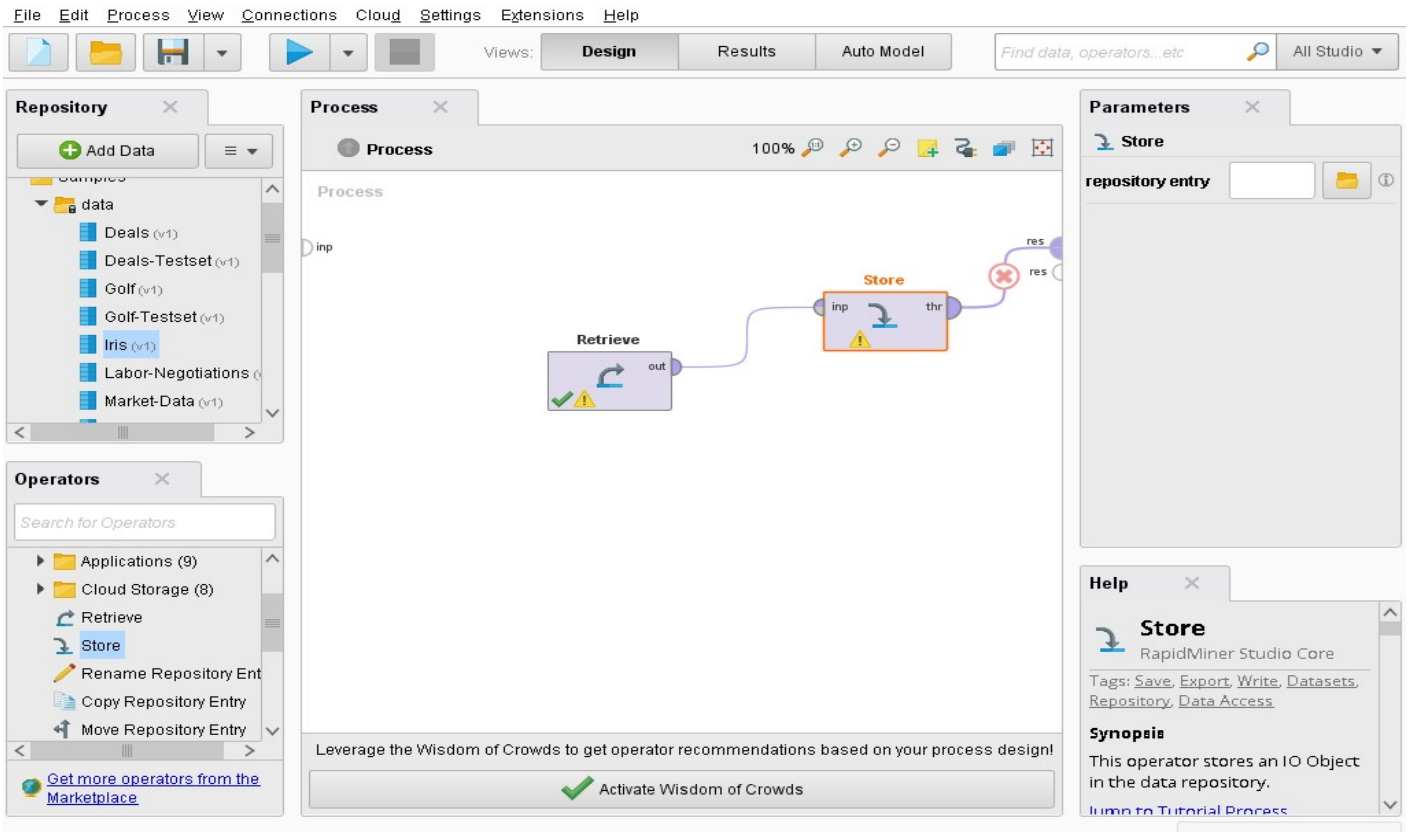
Step-10 Start Execution of Current Process



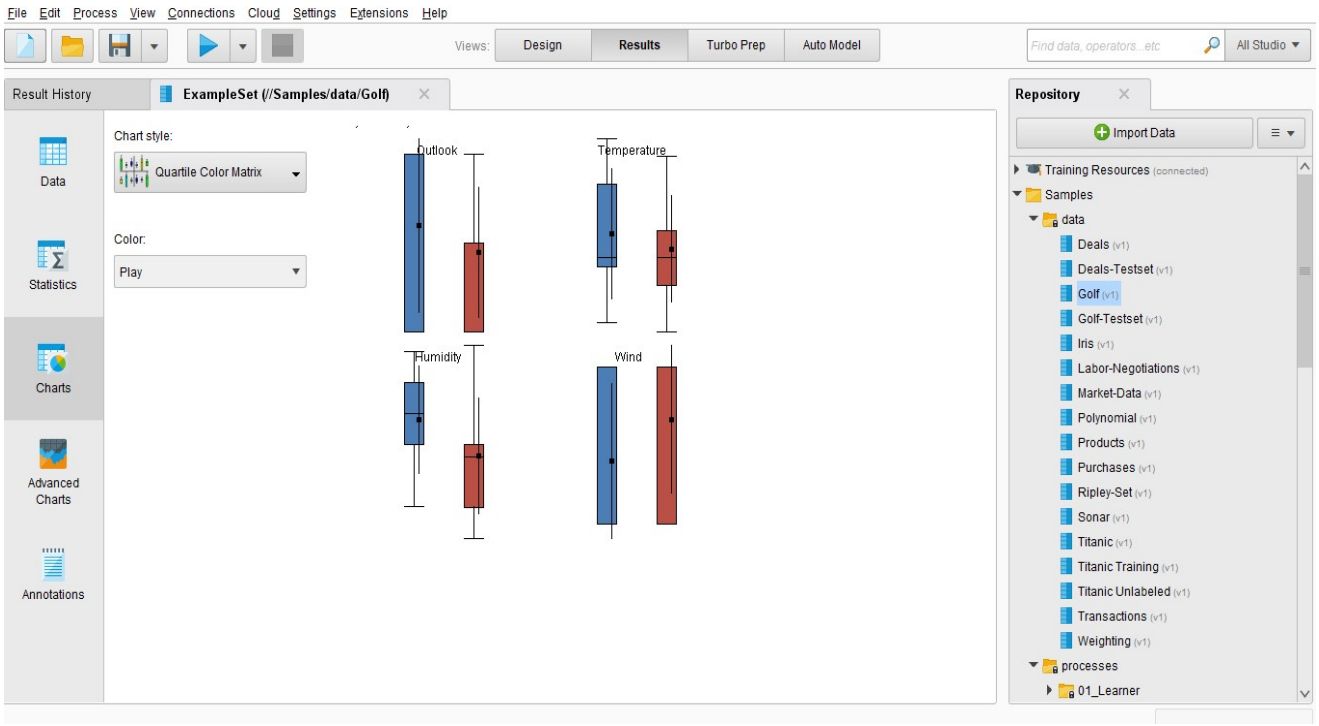
Step-11 Output Result Generated after Execution of Currrent Process

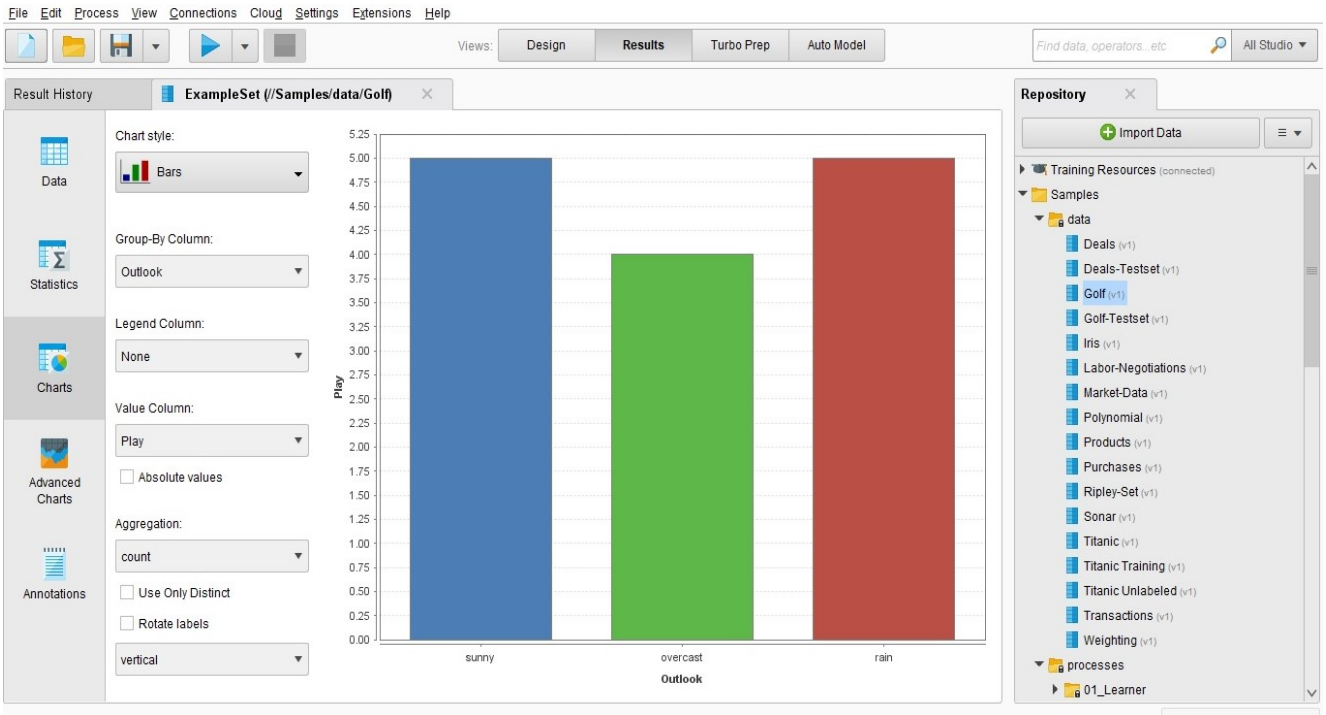


Step-12 Now you can add Store Operator and connect to result operator



Step-13 You can also plot Charts of Sample Data set





1.9 Conclusion

With the help Rapid Miner Tool we can Perform ETL operations on Sample Data sets and can perform analysis on sample data sets.

References:-

1. https://career.guru99.com/top-18-data-analyst-interview-questions/
2. https://docs.rapidminer.com/downloads/RapidMiner-v6-user-manual.pdf